How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation

Year: 2023 | Citations: 531 | Authors: Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr

---

Abstract

Generative Pre-trained Transformer (GPT) models have shown remarkable capabilities for natural language generation, but their performance for machine translation has not been thoroughly investigated. In this paper, we present a comprehensive evaluation of GPT models for machine translation, covering various aspects such as quality of different GPT models in comparison with state-of-the-art research and commercial systems, effect of prompting strategies, robustness towards domain shifts and document-level translation. We experiment with eighteen different translation directions involving high and low resource languages, as well as non English-centric translations, and evaluate the performance of three GPT models: ChatGPT, GPT3.5 (text-davinci-003), and text-davinci-002. Our results show that GPT models achieve very competitive translation quality for high resource languages, while having limited capabilities for low resource languages. We also show that hybrid approaches, which combine GPT models with other translation systems, can further enhance the translation quality. We perform comprehensive analysis and human evaluation to further understand the characteristics of GPT translations. We hope that our paper provides valuable insights for researchers and practitioners in the field and helps to better understand the potential and limitations of GPT models for translation.